

In 2014, further samples from Te Hauturu-o-Toi/Little Barrier Island, Otata and the Kermadec Islands were collected, and genetic material (from microorganisms in the soil through to DNA shed by animals aboveground) was extracted and sequenced by PhD student Andrew Dopheide and technician Laura Ward. The result is a total of many millions of DNA sequences for analysis. The quantity of data generated is vast as this project is considering thousands of species rather than a handful.

The year has predominantly focused on developing efficient ways to analyse the enormous amounts of genomic, environmental and taxonomic information. The pipeline developed by Dr Walter Xie in Professor Alexei Drummond's team at the University of Auckland has been customised to ensure a smooth transition between the different software packages used. The focus was on genes from bacteria, fungi and invertebrates, and on examining biodiversity on an ecological level (eg determining that bacteria at lower altitudes tend to be adapted to low pH soil conditions).

The initial data are correlating well with traditional methods of biological assessment - particularly for invertebrates, which matched well for the different altitudes sampled on Little Barrier Island. A paper was submitted to *GigaScience*, with publication expected in 2015.

The refined process means that analysis should be more rapid in 2015, so the groundwork is laid for exciting insights into the drivers of diversity within ecosystems, finding novel species, and looking for groups of taxa that are indicative of a healthy ecosystem. 2015 will also focus on improving methods for amplifying plant and vertebrate DNA from soil samples.

The collection, sequencing and analysis of the samples is supported by Victoria University of Wellington, The University of Auckland, Landcare Research, and Plant & Food Research.

Collaboration has led to significant national impact. *Hidden Treasures* is the case study for a science teaching resource that introduces pupils to the importance of ecosystems and conservation. Work on comparing methodologies for assessing biodiversity, with a view to standardising methods (eg using the same genes, substrate collection, DNA extraction, and analysis), has been incorporated into the Biological Heritage National Science Challenge and AWC researchers were involved in writing the Challenge's research plan.

The AWC was also a founding member of the Genomics Observatories Network, launched in 2014 by 57 research groups from international organisations including museums (eg Australian Museum, Smithsonian National Museum of Natural History), universities (eg Universities of Auckland, Oxford, Berkeley), and research institutions (eg Max Planck Institute for Marine Microbiology, British Antarctic Survey).

Imaging Evolution: Mathematics of life

Exploring possible evolutionary scenarios for any living organism requires the mathematical handling of huge volumes of data. It's not enough, or even remotely accurate in many cases, to assume that species have evolved from a common ancestor by a simple branching process.

The Imaging Evolution strategic initiative, co-ordinated by Allan Wilson Centre Deputy Director Professor Mike



Steel, has developed several strands over the past year that have seen novel mathematics and software developed to help biologists discover more about the evolution of life.

PopART

One strand with increasing national and international uptake is the software PopART, aka Population Analysis with Reticulate Trees, to help biologists understand evolutionary relationships among populations. The software was developed, using new mathematical algorithms, to show the complex nature of reticulate evolution using haplotype networks where species, particularly bacteria, can develop through hybridisation between species or by horizontal gene transfer (the transfer of genes between organisms in ways other than reproduction). The software can also link genetic data with ecological, geographical and demographic information.

The free population genetics software was developed by mathematicians and biologists from five universities and research institutes across New Zealand. The primary developer is Dr Jessica Leigh. The other main contributors are Associate Professor David Bryant and Mike, as well as a dedicated group of beta-testers.

"The idea is that biologists can get genetic sequences from a population and construct a network that reveals the population's migration patterns or the way that the population has dispersed over time, and how different subpopulations are related to one another. So, it can be used for analysing within a species rather than between species," says Mike.

"Part of the motivation for this direction with haplotype networks is that the software packages that biologists have been using were woefully inadequate and poorly justified. Jessica and David decided that we needed to start from scratch and design a mathematically clear way of representing these networks."

Jessica says there are several reasons why PopART is beginning to stand out among other population genetics software. "Firstly, because it's still in development I'm constantly adding features requested by the users. My goal is to gradually build a full-feature platform that can do all of the basic things that most population geneticists need their software to do, and that runs on all major operating systems, such as Windows, Mac OSX and Linux".

"Our goal with PopART is to provide a single piece of software that performs most of the analyses that most people need, and that produces publication-ready figures, including map figures that display geographical distribution of sequence data. It's still a work in progress, but I've gotten a lot of positive feedback so far," says Jessica.

Mike agrees, "I understand from having been in touch with a couple of Allan Wilson Centre groups using the software that it's working well. For example, Nigel French's group in Palmerston North has been using PopART with analysis research into pathogen evolution, things like *Campylobacter* and various other nasties."

Jessica believes there are over 100 research groups now using PopART, with the number growing as word spreads, "Most PopART users now aren't in New Zealand. At one point, I think most international users had some connection to New Zealand—there are groups of users in Sweden and Switzerland who have connections to former AWC students. But now I think they're a few degrees removed from New Zealand researchers."

Mathematics for modelling lateral gene transfer

Another Imaging Evolution strand being concurrently developed is new mathematics for modelling lateral gene transfer. Lateral gene transfer is the main cause of bacteria evolving antibiotic and pesticide resistance. It's been found that this type of gene transfer can occur between species and is the mechanism for conferring resistance to other species of bacteria - creating one of the biggest medical challenges for the 21st Century.

Mike explains his team's recent work in modelling horizontal or lateral gene transfer, "We've been doing some work with biologists in the United States to come up with a way of modelling lateral gene transfer. This type of transfer tends to be more prevalent in bacteria and microorganisms which traditionally just reproduce asexually by splitting, but what happens sometimes is they get together and exchange genetic material - genes pass between them, giving an organism new genes within its lifetime that it can pass on to offspring.

"There is a lot of debate about how extensive this type of transfer is, it is believed to have been more common in the early days when life began, but it is still a process that occurs today. Of course, this type of gene transfer mucks up the whole story of building an evolutionary tree and there has been a lot of debate about should we just throw the tree model out. Well, I think most biologists would say that the tree is still a useful model for describing eukaryotes, which are fungi, plants and animals—most of the things we know and love. However, amongst microorganisms lateral gene transfer is a real issue, so we've come up with a mathematical analysis of that, which we published last year. We were able to reconstruct networks and trees based on four bacteria species that showed very high levels of lateral gene transfer."

Mike has presented and discussed the findings at several international conferences and overseas institutions since the papers were published.

The origins of life

A third strand being undertaken by the Imaging Evolution strategic initiative is new mathematical approaches for network methods in other areas of evolution. Mike says that the past two years has seen them turn their attention to the origins of life itself.

"The origin of life nearly four billion years ago is a very complex and difficult question and there is a lot of debate about how life could have started - it's the sort of field where there is lots of opinion but very little data. We may never know exactly how it started, but a good scientific objective would be to describe how it might have happened - that is, what is a plausible scenario?"

“So, in discussion with biologists, some of whom have various hypotheses, we realised that you could take very complicated metabolic systems, for example, in *E. coli* bacteria, and systematically search and figure out what subsystems hidden within there might have been able to be self-sustaining what’s called autocatalytic.”

Autocatalysis, where each reaction is catalysed by the product of some other reaction, is widely believed to be one of the requirements of the transition from just chemistry to living cells in the Earth’s primordial soup.

“All biochemical reactions in a cell are catalysed - they have to happen quickly and they have to happen in a synchronised fashion. We had formalised some time ago this notion of a reflexively autocatalytic and food-generated (RAF) sets. We had this algorithm which could quickly identify RAFs.”

“One of our former AWC post docs Wim Hordijk found a paper published in *Nature* about a year ago, where a research group claimed to have produced the first very basic self-sustaining ‘living’ system of mutually autocatalytic ribozymes in the lab. We were able to reanalyse all their data using our methods and pick up new features that they completely missed in their analysis, so that was quite exciting. This resulted in another paper.”

“But, what we really wanted to do this year was take one of the most primitive organisms living on Earth today, *Methanopyrus kandleri*, it’s what is called an *Archaea*, a single-celled organism living on the sides of hot hydrothermal vents in the deep ocean. It has relatively few genes and a very simple structure, with no organelles. However, the problem with that was not enough is known

about its biochemistry, so we had to go for *E. coli*. By comparison, *E. coli* is hugely complex, nevertheless, it’s been so well studied that its entire metabolism is known, what catalyses what is known. So, a post doc in Germany, Filipa Sousa, working with her supervisor Professor Bill Martin, and computational biologist Wim Hordijk, fed all these thousands of reactions and molecules into our system to identify these autocatalytic sets.”

The research team found that the largest RAF encompassed almost the entire *E. coli* cytosolic reaction network. They systematically studied its structure by removing catalysts or reactions, to find the minimum food set that maintains a given RAF. They found that RAFs capture properties of biological organization at the level of single cells, with the interdependency between different metabolic modules pointing to the important role of spontaneous (non-enzymatic) reactions in early evolution.

Mike says that while the work gave them new insights into the metabolism of *E. coli*, the main motivation wasn’t to understand *E. coli*, but to see what they could learn about early life.

“What we’ve managed to do is provide one of the most promising mathematical foundations for studying early life. The leading figures working in the field are now routinely citing this work. One of its advantages was we started with minimal assumptions, it meant we could find things quickly without getting bogged down. And we’ve recently heard that our paper on the project has been accepted for publication in the *Journal of Systems Chemistry*.”

AWC workshop

As well as AWC Imaging Evolution members presenting at numerous workshops and conferences around the world in the past year, Mike says they also organised a workshop of their own. “We had a Networks of Life workshop in Kaikoura in February. It was a week-long meeting that brought together graduate students and overseas experts to talk about new ideas, discuss new research, and carry-out complicated equations on the whiteboard. It was a very successful meeting and resulted in some new ideas, new research collaborations, and new papers being published. We’re now planning a further workshop in February 2015.”

Two papers which Mike says sprung from the workshop were *Tree-like Reticulation Networks-When Do Tree-like Distances Also Support Reticulate Evolution?* by Andrew Francis (University of Western Sydney) and Mike Steel published in *Mathematical Biosciences* in January 2015; and *The agreement problem for unrooted phylogenetic trees is FPT* by Celine Scornavacca (University of Montpellier II, France), Leo van Iersel and Steven Kelk (both from The Netherlands) and David Bryant published in the *Journal of Graph Algorithms and Applications* in June 2014.



Jessica Leigh, the primary developer of PopART, is a former AWC postdoctoral fellow in the Department of Mathematics and Statistics at the University of Otago, working with Associate Professor David Bryant. Following a move to Auckland she remains on contract with the AWC. Her current research interest is in inference and visualisation of haplotype networks; and a Bayesian framework for simulation experiments.

Dr Jessica Leigh. Credit: Steffen Klaere.